bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Genetic demultiplexing of pooled single-cell RNA-sequencing samples in cancer facilitates effective experimental design

Lukas M. Weber[1], Ariel A. Hippen[2], Peter F. Hickey[3], Kristofer C. Berrett[4], Jason Gertz[4], Jennifer Anne Doherty[4], Casey S. Greene[2], Stephanie C. Hicks[1]*

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[2] Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, PA, USA
[3] Advanced Technology & Biology Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia
[4] Huntsman Cancer Institute and Department of Population Health Sciences University of Utah, UT, USA

* Corresponding author

6 November 2020

Keywords: genetic demultiplexing, single-cell RNA sequencing, cancer, high-grade serous ovarian cancer, lung adenocarcinoma, computational methods, simulations, benchmarking

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Pooling cells from multiple biological samples prior to library preparation within the same single-cell RNA sequencing experiment provides several advantages, including lower library preparation costs and reduced unwanted technological variation, such as batch effects. Computational demultiplexing tools based on natural genetic variation between individuals provide a simple approach to demultiplex samples, which does not require complex additional experimental procedures. However, these tools have not been evaluated in cancer, where somatic variants, which could differ between cells from the same sample, may obscure the natural genetic variation. Here, we performed *in silico* benchmark evaluations by combining sequencing reads from multiple single-cell samples in high-grade serous ovarian cancer, which has a high copy number burden, and lung adenocarcinoma, which has a high tumor mutational burden, confirming that genetic demultiplexing tools can be effectively deployed on cancer tissue using a pooled experimental design. We demonstrate that this strategy provides significant cost savings through pooled library preparation. To facilitate similar analyses at the experimental design phase, we provide freely accessible code and a reproducible Snakemake workflow built around the best-performing tools found in our *in silico* benchmark evaluations, available at https://github.com/lmweber/snp-dmx-cancer.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Sample pooling prior to library preparation is an effective strategy for experimental design in single-cell RNA sequencing (scRNA-seq) studies, which allows researchers to assess and address unwanted technological variation such as batch effects [1,2] and reduces library preparation costs [3–5]. Several strategies involve pooling cells, labeled or otherwise identifiable in some way, from multiple biological samples, followed by combined library preparation and sequencing, and computational demultiplexing to recover the sample identities of each cell. While sample pooling has been shown to lead to increased doublet rates [5], demultiplexing approaches can identify doublets at the demultiplexing step without relying on downstream doublet identification tools [6–10]. Depending on the method used, these techniques can also avoid the phenomenon of sample index swapping, which occurs when individually prepared libraries are subsequently pooled for sequencing [11–15].

Existing demultiplexing approaches differ in their experimental procedures, computational methodology for demultiplexing, and demultiplexing accuracy. In barcoding-based approaches (e.g. MULTI-seq [16] and cell hashing [17], and GMM-Demux for doublet identification [18]), cells are experimentally tagged with universal oligonucleotides or antibodies together with sample-specific labels, which can give highly accurate demultiplexing performance, but these approaches make sample preparation more complex and increase costs due to reagent purchases as well as additional library preparation and sequencing. Alternatively, genetic variation-based approaches rely only on natural genetic variation between samples from different individuals (such as single nucleotide polymorphisms, SNPs), which does not require additional experimental procedures at the single-cell level. Initial genetic variation-based demultiplexing methods, such as demuxlet [5], require a known genotype reference for each sample obtained using SNP arrays, whole exome sequencing, or bulk RNA sequencing. Recently, new methods have been developed, such as Vireo [3], scSplit [4], souporcell [19], and freemuxlet [20], which can use probabilistic models to infer the genotype directly from the single-cell reads. Depending on the method, there is also the option to improve performance by providing either sample-specific genotypes, such as from matched bulk RNA sequencing, or a list of population SNPs, such as from the 1000 Genomes Project [21] for human samples.

However, genetic variation-based scRNA-seq demultiplexing tools have not yet been evaluated in cancer [3–5]. Cancer is characterized by widespread additional somatic mutations, including single nucleotide variants (SNVs) [22], as well as structural variation affecting the frequency of SNVs, which could interfere with the SNP signal used to distinguish individuals for demultiplexing. The frequency of additional somatic SNVs, known as the tumor mutational burden (TMB), can vary widely between cancer types [23], as well as between patients and cancer subtypes [24,25]. However, the TMB is typically small relative to the overall population SNP burden [22]. For example, population SNPs with minor allele frequency (MAF) >1% are thought to occur on the order of once per 1000 nucleotides on average, or 1000 SNPs per Mb [26]. By contrast, high-TMB cancers have been defined as having around >10 or >20 additional mutations (SNVs) per Mb [24,25] -- approximately two orders of magnitude lower frequency than the population SNPs. Therefore, it seems reasonable to expect that the natural genetic variation signal would not be severely obscured by the TMB, and that genetic variation-based demultiplexing tools should still perform well in cancer. However, this assumption has not been tested. Due to the finite and irreplaceable nature of tumor samples, we computationally evaluated demultiplexing algorithms to confirm that genetic variation-based demultiplexing performs adequately when applied to scRNA-seq pooling experimental designs in cancer, before committing samples to this experimental

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

design strategy. In addition, we were interested in evaluating the degree to which these tools can reliably identify doublets, including in experimental designs with relatively high proportions of doublets. Reliable doublet identification would allow the use of "super-loading" experimental designs, i.e. loading cells at high concentration and subsequently removing identifiable doublets consisting of cells from multiple samples, providing substantial cost savings during library preparation [5,17,27]. In the future, these tools may also be well-suited for cell atlas initiatives, which are expected to cover large numbers of samples, including eventually those from cancer [28,29].

Here, we performed a benchmark evaluation of genetic variation-based demultiplexing in cancer scRNA-seq data using *in silico* simulations constructed from experimental scRNA-seq datasets with known sample identity for each cell. We evaluated two demultiplexing algorithms and five strategies for identifying the genotype reference list of SNPs used in the demultiplexing algorithms. In addition, we included varying proportions of simulated doublets by combining sequencing reads from multiple cell barcodes. In the benchmark evaluation, we considered scRNA-seq samples from cancers that should be difficult to characterize: high-grade serous ovarian cancer (HGSOC) and lung adenocarcinoma. HGSOC is characterized by loss of TP53, and generally has medium to high SNV burden and high copy number variation (CNV) burden (particularly for focal copy number alterations), relative to other cancers [23,30], while lung adenocarcinoma is characterized by high SNV burden [23]. Our results demonstrate that genetic variation-based demultiplexing provides high recall at acceptable precision-recall tradeoffs in both high CNV and high SNV cancer types, even with high simulated doublet proportions. Our results demonstrate that these tools support experimental designs that incorporate sample pooling. We provide a reproducible Snakemake [31] workflow based on the best-performing combination of tools for estimating a genotype reference list of SNPs and demultiplexing samples identified in our benchmark, to facilitate experimental design efforts. The Snakemake workflow is modular, allowing users to substitute alternative tools. The workflow requires a set of scRNA-seq pilot samples, access to a Linux computing cluster, some familiarity with the Linux command line, and optionally matched bulk RNA-seq samples (for the highest demultiplexing performance and doublet identification). All code for the benchmark evaluation and Snakemake workflow is freely accessible at https://github.com/lmweber/snp-dmx-cancer.

Results

Genetic demultiplexing in HGSOC and lung adenocarcinoma

We evaluated the performance of genetic demultiplexing algorithms for scRNA-seq samples from HGSOC (high CNV) and lung adenocarcinoma (high SNV) using a set of benchmark evaluations and Snakemake workflow built around freely available tools including Cell Ranger [32], samtools [33], bcftools [34], Unix string manipulation tools (sed and awk), cellSNP [35], and Vireo [3] (**Methods** and **Figure 1**). The HGSOC samples were collected at the Huntsman Cancer Institute, and the lung adenocarcinoma dataset is a published dataset sourced from [36]. **Table 1** provides a summary of the scRNA-seq datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

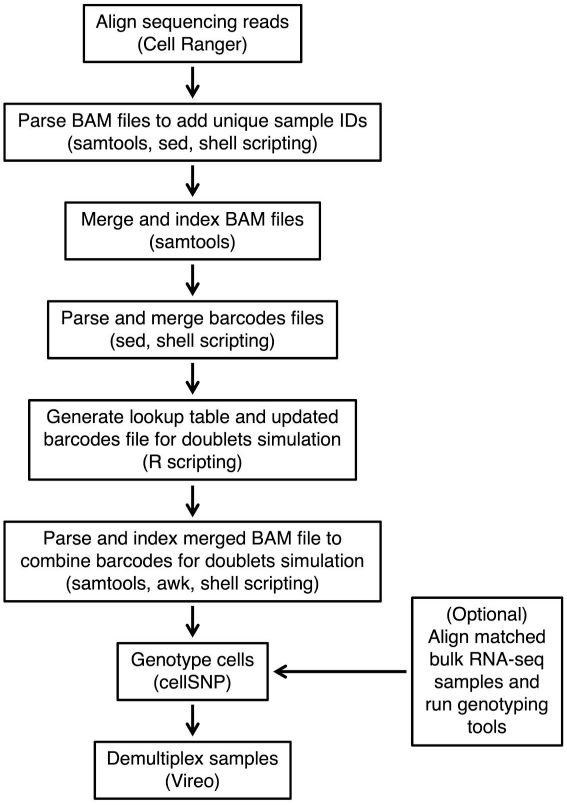

Figure 1. Schematic illustrating the steps in the Snakemake workflow. The workflow is designed to be modular, allowing users to substitute alternative tools. The Snakemake workflow runs a complete analysis for one dataset (HGSOC) and doublets simulation scenario (20% doublets). Our benchmark evaluations include a second dataset (lung adenocarcinoma) and additional doublet simulation scenarios (30% doublets, no doublets). The optional step to run genotyping tools (e.g. on matched bulk RNA-seq samples) improved performance in our benchmark. Tools used in each step are shown in parentheses.

HGSOC dataset		**Lung adenocarcinoma dataset**	
Sample ID	Number of cells	Sample ID	Number of cells
X2	7123	T08	4093
X3	1533	T09	4267
X4	6546	T20	4521
		T25	4428
		T28	5789
		T31	7069

Table 1. Summary of number of samples and number of cells per sample for scRNA-seq samples in HGSOC (GSE158937 and phs002262.v1.p1) and lung adenocarcinoma [36] (EGAD00001005054) datasets. The numbers of cells per sample listed are the numbers of cells provided by Cell Ranger [32] following sequencing read alignment. The HGSOC dataset additionally includes matched bulk RNA-seq samples for each sample. The lung adenocarcinoma dataset includes matched bulk whole exome sequencing samples for each sample, but not matched bulk RNA-seq samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

High precision and recall performance using genetic demultiplexing

Using the HGSOC scRNA-seq and matched bulk RNA-seq data, we found the highest recall (defined as the proportion of true singlet cells for each sample that are identified as singlets and assigned to the correct sample) and best precision-recall tradeoff (where precision is defined as the proportion of identified cells for each sample that are true singlet cells from the correct sample), when using bcftools [34] to generate a genotype reference list of SNPs from the matched bulk RNA-seq samples, together with cellSNP/Vireo [3,35] for demultiplexing, in all simulation scenarios (30% doublets, 20% doublets, or no doublets) (**Figure 2 a-c**). This scenario (labeled "bulkBcftools_cellSNPVireo" and colored light blue in **Figure 2**) achieves 99.9%, 99.9%, and 99.0% recall (values averaged across three scRNA-seq samples). However, in this scenario, the precision drops (77.4%, 85.9%, and 100%) (values averaged across three scRNA-seq samples) as the percentage of doublets increases with 30%, 20%, and no doublets (**Figure 2 a-c**, panels from left to right), respectively. In general, we prefer higher recall at the expense of somewhat lower precision, so that we do not lose informative singlet cells, and since additional doublet detection tools [6–10] can potentially be applied during downstream analyses to further improve precision.

We also found that using bcftools [34] to generate a genotype reference list of SNPs from the matched bulk RNA-seq samples together with demuxlet (labeled "bulkBcftools_demuxlet" and colored green in **Figure 2**) resulted in somewhat higher precision (84.3%, 91.3%) with a large reduction to recall (52.1%, 53.0%) in the 30% and 20% doublet scenarios, respectively (**Figure 2 a-b**). However, no further improvement in precision was observed (99.9%) with a large reduction in recall (52.8%) for the no doublets scenario (**Figure 2 c**).

In the scenarios where matched bulk RNA-seq samples are not available, the next best-performing scenarios were obtained using the genotype reference from the 1000 Genomes Project [21] (provided by the authors of cellSNP/Vireo) with no filtering of SNPs ("unfiltered") and genotype reference from the 1000 Genomes Project filtered to retain only SNPs in the 3' untranslated region (UTR) ("filtered"), together with cellSNP/Vireo for demultiplexing (labeled "1000GenomesUnfilt_cellSNPVireo" and "1000GenomesFilt_cellSNPVireo" and colored in orange and purple, respectively) (**Figure 2 a-c**). The "unfiltered" scenario achieved recall 99.0%, 99.0%, and 97.9%, and precision 74.5%, 84.0%, and 100%, with 30% doublets, 20% doublets, and no doublets respectively (**Figure 2 a-c**, panels from left to right). The "filtered" scenario achieved recall 95.3%, 95.0%, and 94.3%, and precision 73.8%, 83.7%, and 100%, respectively. Alternatively, when we evaluated the scenario to call SNPs directly from the scRNA-seq samples and use cellSNP/Vireo for demultiplexing (labeled "singlecellCellSNP_cellSNPVireo" and colored in dark blue), we found comparable recall (92.1%, 91.9%, 91.5%) with a slight loss in precision (72.7%, 82.3%, 99.7%) as the percentage of doublets increases with 30%, 20%, and no doublets, respectively (**Figure 2 a-c**, panels from left to right).

Using the high-TMB lung adenocarcinoma scRNA-seq (without matched bulk RNA-seq) dataset, we only considered the scenario using the genotype reference from the 1000 Genomes Project (filtered) together with cellSNP/Vireo (labeled "1000GenomesFilt_cellSNPVireo" and colored in purple in **Figure 2**), as this resulted in the highest precision and recall in the HGSOC evaluation when using either the genotype reference from 1000 Genomes Project or directly calling SNPs from the scRNA-seq samples, while also keeping runtimes much lower than the 1000 Genomes (unfiltered) scenario. In this scenario (labeled "1000GenomesFilt_cellSNPVireo"), we found comparable ranges of precision and recall values as for the matching scenario in the HGSOC dataset (**Figure 2 d-f**). These results demonstrate that we can also achieve excellent demultiplexing performance even in a higher-TMB cancer setting.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

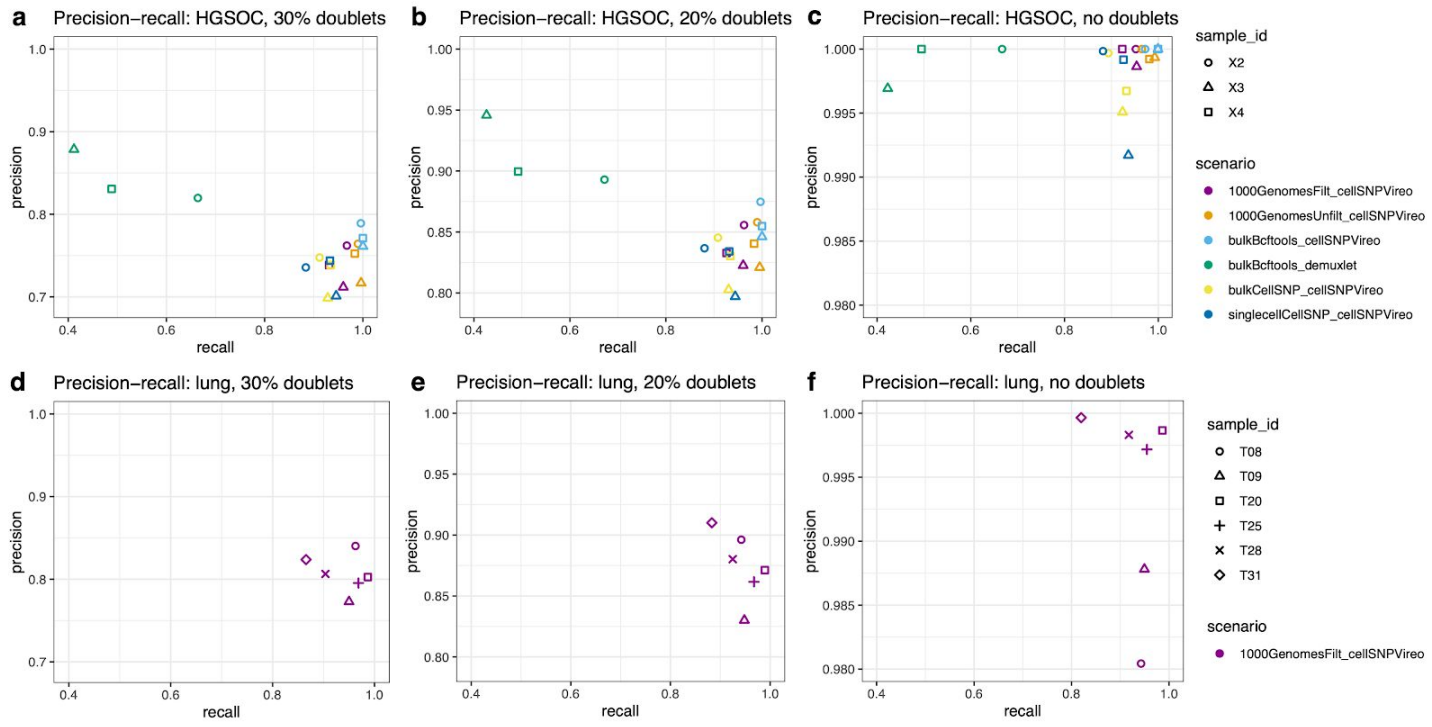

Figure 2. Performance evaluations for benchmark scenarios, for HGSOC dataset **(a-c)** and lung adenocarcinoma dataset **(d-f)**, across three proportions of simulated doublets (30%, 20%, and no doublets). Performance is evaluated in terms of precision (y-axis) and recall (x-axis) for recovering the sample identities of singlet cells for each scRNA-seq sample. Benchmark scenarios are labeled by color and with the naming scheme "genotypeMethod_demultiplexingMethod". Samples within each dataset are identified with shapes. Note that y-axis limits (precision) are the same between rows of panels (datasets), but differ between columns of panels (doublet proportions) for improved visibility.

Computational runtime of genetic demultiplexing workflow steps and genotyping tools

We evaluated the computational runtimes for the various components in our benchmark scenarios and Snakemake workflow using the HGSOC data. First, we found the computational runtimes for the various steps in the genetic demultiplexing workflow vary across multiple orders of magnitude and depended on whether the tool could be parallelized. The parallelizable tools (Cell Ranger and cellSNP) were run using 10 processor cores to decrease runtime, while the remaining tools used a single core. All evaluations of runtimes were performed on a high-performance Linux computing cluster. In the Snakemake workflow (**Figure 3 a**), the slowest steps were running Cell Ranger (approximately 6 hours per sample using 10 cores) and cellSNP (approximately 5 hours using 10 cores), and parsing the merged BAM file containing aligned reads to combine cell barcodes into simulated doublets (approximately 1 day). For the cellSNP step in the workflow, runtime depended on the choice of genotype reference list of SNPs (**Figure 3 b**). In particular, filtering the genotype reference from the 1000 Genomes Project [21] (provided by the authors of cellSNP/Vireo) to retain only SNPs in the 3' UTR reduced runtime from approximately 3 days to less than 1 hour ("1000GenomesUnfilt_cellSNP" vs. "1000GenomesFilt_cellSNP"), at the cost of only a small drop in performance (**Figure 2**). The runtime shown for the cellSNP step in **Figure 3 a** corresponds to the highest-performing scenario from **Figure 2** ("bulkBcftools_cellSNP").

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We also evaluated computational runtimes for the genotyping tools used to generate the genotype reference lists of SNPs from either the matched bulk RNA-seq samples or directly from the scRNA-seq samples (**Figure 3 c**). Here, we found by far the slowest option was to use cellSNP to generate the genotype reference directly from the scRNA-seq samples (between 1 and 4.5 days per sample using 10 cores), while generating the genotype reference from the bulk RNA-seq samples took either 3-5 hours per sample using cellSNP (10 cores) or 5 hours using bcftools.

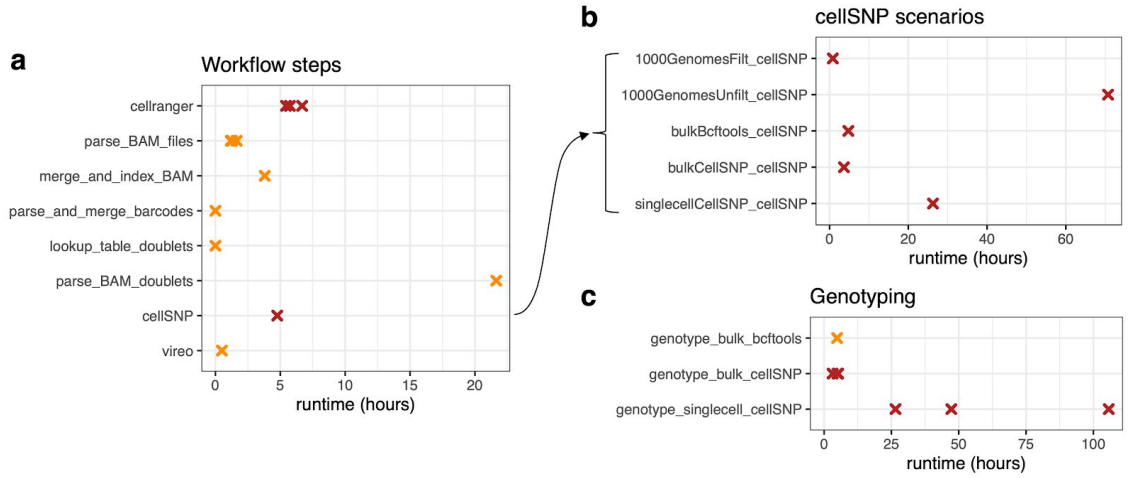

Figure 3. Computational runtimes (in hours) of genetic demultiplexing workflow steps and genotyping tools. **(a)** Runtimes for steps in the complete Snakemake workflow, for a single dataset (HGSOC) and doublets simulation scenario (20%). Parallelized tools (Cell Ranger and cellSNP; points indicated in dark red) were run using 10 processor cores, and all other tools using a single core (points indicated in orange), on a high-performance Linux computing cluster. For steps where samples were processed individually, separate points are shown for each sample. **(b)** Runtimes for alternative options for running cellSNP in the workflow, depending on the choice of genotype reference (1000 Genomes filtered, 1000 Genomes unfiltered, matched bulk RNA-seq using bcftools, matched bulk RNA-seq using cellSNP, and single-cell RNA-seq using cellSNP). The cellSNP step in (a) matches the row “bulkBcftools_cellSNP” in (b), which was the highest-performing scenario from Figure 2. **(c)** Runtimes for alternative options to generate genotype reference file. Note that horizontal axis scales differ between panels for improved visibility.

Estimated cost savings due to multiplexing

As an illustration of expected cost savings due to lower library preparation costs in a multiplexed experimental design, we estimated library preparation and sequencing costs for experimental designs with 4 to 8 samples, using the “Cost Per Cell” online calculator provided by the Satija Lab [27] (**Figure 4**). For the calculations, we assumed 4,000 desired cells per sample after demultiplexing, i.e. after discarding identifiable doublets consisting of cells from multiple samples, but including the smaller number of non-identifiable doublets (multiple cells from the same sample, which have the same SNP profiles and cannot be distinguished using genetic demultiplexing). We used default settings for the remaining assumptions, including library preparation costs of $2,000 per sample or multiplexed set of samples; sequencing costs of $1,500 per 400 million reads with an additional 30% cost due to unaligned reads and adapters; and approximately 20,000 reads per cell.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4 shows the combined library preparation and sequencing cost with either no multiplexing or full multiplexing (all samples prepared as a single library and sequenced together) in each case.

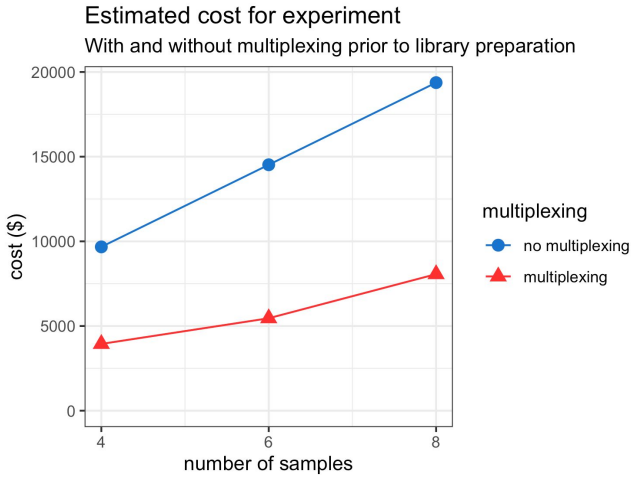

Figure 4. Illustration of expected cost savings from multiplexed experimental design prior to library preparation. The figure shows the total of estimated library preparation and sequencing costs, with either no multiplexing or full multiplexing (all samples prepared as a single library and sequenced together), for experiments with 4, 6, or 8 samples. The calculations assume 4,000 desired cells per sample after demultiplexing, after discarding identifiable doublets consisting of cells from multiple samples; library preparation costs of $2,000 per sample or multiplexed set of samples; sequencing costs of $1,500 per 400 million reads with an additional 30% cost due to unaligned reads and adapters; and approximately 20,000 reads per cell. Calculations were performed using the “Cost Per Cell” online calculator provided by the Satija Lab [27].

Discussion

Pooled single-cell experimental designs before library preparation together with genetic variation-based computational sample demultiplexing are a convenient and effective strategy for reducing library preparation costs and potential batch effects in scRNA-seq studies. Here, we performed an *in silico* benchmark evaluation using real scRNA-seq datasets to confirm that these tools can be effectively applied in cancer data. We selected HGSOC and lung adenocarcinoma, cancer types characterized by a relatively high TMB. Previously, these tools have only been evaluated in non-cancer datasets, which are not affected by additional mutational SNV burden that could potentially obscure the natural genetic variation SNP signal used to distinguish individuals. Our benchmark evaluations include simulated proportions of doublets (up to 30%), confirming that these tools can be used to identify singlet cells in “super-loading” experimental designs to achieve considerable cost savings in library preparation [5,17,27]. In our HGSOC dataset, we achieved the best demultiplexing performance (and relatively efficient runtimes) when using matched bulk RNA-seq samples to generate a genotype reference list of SNPs using bcftools [34], together with cellSNP/Vireo [3,35] for demultiplexing. However, using a standard list of population SNPs from the 1000 Genomes Project [21] (which does not require matched bulk RNA-seq samples) provided by the authors of cellSNP/Vireo also achieved good performance. In this case, filtering the population SNPs to retain only SNPs in the 3’ UTR significantly reduced runtime, at the cost of only slightly lower demultiplexing performance. For the lung adenocarcinoma dataset, performance was comparable to the matching scenario in the HGSOC dataset, confirming that

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

performance was not seriously affected by the higher TMB, and that genetic demultiplexing can be effectively applied in this setting. Since most other cancer types have lower TMB [23], we expect these results to apply to most cancer types. We provide a freely available, modular Snakemake [31] workflow implementing the best-performing scenario from our benchmark evaluations, built around cellSNP/Vireo [3,35] and other freely accessible tools, as well as additional R and shell scripts to reproduce all analyses in our benchmark evaluation (https://github.com/lmweber/snp-dmx-cancer), to allow other researchers to perform similar analyses for experimental design, planning, and budgeting purposes in their own datasets.

Our study has several limitations. While the best-performing benchmark scenario achieves excellent recall, precision is somewhat lower. This could be addressed using additional downstream tools to identify doublets [6–10]. In general, we prefer higher recall at the expense of somewhat lower precision, so that we are not losing informative cells during the initial steps of the analysis. In this study, we have built our evaluations around the best-performing tools (cellSNP/Vireo [3,35] for demultiplexing and using matched bulk RNA-seq samples for genotyping) and compared against alternative tools, such as demuxlet [5] and baseline scenarios (no doublets), but we have not performed a comprehensive benchmark evaluation of all available tools, such as additional tools for demultiplexing (e.g. scSplit [4], souporcell [19], and freemuxlet [20]). However, we have implemented the Snakemake workflow to be modular, so that other users may substitute alternative tools if they prefer. We also investigated the use of salmon alevin [37] for pseudoalignment of scRNA-seq reads (instead of Cell Ranger), but found that this was not compatible with the demultiplexing tools since pseudoalignment occurs at the transcriptomic instead of genomic level. However, future developments may enable conversion between transcriptomic and genomic aligned reads, and we have included alternative code scripts for salmon alevin within our benchmark code repository. Our evaluations only considered two tumor types (HGSOC and lung adenocarcinoma), and performance may differ for other cancer types or tissues. However, since we were able to demonstrate good performance in lung adenocarcinoma, one of the highest TMB cancers, we anticipate these results will also be applicable for other cancer types, which will generally have lower TMB. For the lung adenocarcinoma dataset, matched bulk whole exome sequencing data were also available for these six samples, which could be used to further improve performance using additional genotyping tools. More fundamentally, due to the reliance on genetically distinct SNP profiles, genetic demultiplexing tools are expected to work well for human samples from unrelated individuals, but are not applicable to biological samples from inbred mice or hereditary related human populations, or samples from the same individual [3]. For some experiments, a useful design strategy may also be to combine genetic-based and barcoding-based multiplexing, e.g. multiple treatments on samples from the same individual. Our Snakemake workflow can be used to demultiplex up to approximately 12 pooled samples without a genotype reference (limited by the demultiplexing algorithm Vireo) -- beyond this, the demultiplexing performance of the Vireo algorithm has been shown to decrease [3]. For larger experiments, if matched bulk RNA-seq samples are not available, multiple sample pools could be used, with demultiplexing done separately for each pool [3]. Splitting an experiment across multiple pools and demultiplexing within each pool also represents an opportunity to implement improved experimental designs to reduce batch effects and confounding. Finally, the Snakemake workflow is relatively computationally intensive, and requires access to a high-performance Linux computing cluster or server.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Benchmark evaluations and workflow

We begin by describing in detail our benchmark evaluation framework, and note that our additional Snakemake [31] workflow is built around the combination of tools that resulted in the best performance from the benchmark evaluation. Specifically, the benchmark and workflow make use of several freely available tools, including Cell Ranger [32], samtools [33], bcftools [34], Unix string manipulation tools (sed and awk), cellSNP [35], and Vireo [3]. The Snakemake workflow is designed to be modular, allowing other alternative or new tools to be substituted. All code for the benchmark evaluation and Snakemake workflow is freely available at https://github.com/lmweber/snp-dmx-cancer.

In our benchmark evaluation, we considered two genetic demultiplexing algorithms: (i) Vireo [3] together with cellSNP [35], and (ii) demuxlet [5] as an alternative genetic-based demultiplexing tool. We evaluated five scenarios for obtaining the genotype reference list of SNPs used in the demultiplexing algorithm: (i) list of population SNPs from the 1000 Genomes Project [21] provided by the authors of cellSNP/Vireo; (ii) list of population SNPs from the 1000 Genomes Project with an additional filtering step to retain only SNPs in the 3' untranslated region (UTR) for faster runtime (this strategy is appropriate for 3'-tag sequencing protocols, but could also be adapted for 5'-tag or full-transcript sequencing); (iii) sample genotyping from matched bulk RNA-seq samples using bcftools [34]; (iv) sample genotyping from matched bulk RNA-seq samples using cellSNP [35]; and (v) sample genotyping from scRNA-seq samples using cellSNP [35]. Scenario (ii) was used for both datasets (HGSOC and lung adenocarcinoma), and the remaining scenarios were applied to the HGSOC dataset only. Scenarios (iii) and (iv) require matched bulk RNA-seq samples, while scenarios (i) and (v) have extremely slow runtimes. Specifically, for the HGSOC dataset, we evaluated performance across several combinations of methods for genotyping and demultiplexing (labeled as "genotypeMethod_demultiplexingMethod" in Results). For the lung adenocarcinoma dataset, we used the list of population SNPs from the 1000 Genomes Project provided by the authors of cellSNP/Vireo, filtered to retain only SNPs in the 3' UTR.

We used two datasets for the benchmark evaluations. The first dataset consists of three unique molecular identifier (UMI)-based scRNA-seq HGSOC samples measured on the 10x Genomics platform [38], obtained from separate, unrelated individuals at the Huntsman Cancer Institute at the University of Utah. We also obtained matched bulk RNA-seq samples from the same three individuals for sample genotyping. The raw data is available by controlled access via the Database of Genotypes and Phenotypes (dbGaP) (phs002262.v1.p1), and processed gene count tables are available from the Gene Expression Omnibus (GEO) (GSE158937). The second dataset consists of six UMI-based scRNA-seq higher-TMB lung adenocarcinoma samples measured on the 10x Genomics platform, previously published by [36]. Raw data for all samples in this study are available by controlled access from the European Genome-phenome Archive (EGA) (EGAD00001005054). For our study, we used six samples identified as having TMB >25 mutations / Mb (see [36], Figure 2d and Methods). **Table 1** provides a summary of the scRNA-seq datasets.

Performance was evaluated in terms of precision and recall for demultiplexing each scRNA-seq sample. We also recorded computational runtime for each step in the workflow and benchmark scenarios. Recall is defined as the proportion of true singlet cells for each sample that are identified as singlets and assigned to the correct

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sample. Precision is defined as the proportion of identified cells for each sample that are true singlet cells from the correct sample. Runtime was evaluated by clock time using the Unix `date` command. We used R version 4.0.2 for random number generation and evaluation steps performed in R, and created figures using ggplot2 [39].

For our benchmark evaluation, we developed three *in silico* simulation scenarios for each dataset -- containing either 30% simulated doublets, 20% simulated doublets, or no doublets. Doublets were simulated by combining cell barcode labels from random sets of two cells in the raw sequencing reads mapped using Cell Ranger [32], so that either 30% or 20% of the final barcodes represent doublets. For example, starting with 15,202 original cells in the HGSOC dataset, 3,508 randomly selected cells were combined with 3,508 other cells to create simulated doublets, leaving 11,694 final cell barcodes, of which 3,508 (30%) represent doublets. The 30% doublets scenario represents the upper end of our planned strategy for a “super-loading” experimental design, i.e. loading multiplexed cells at high concentration to reduce library preparation costs and subsequently removing identifiable doublets [5,17,27]; the 20% doublets scenario represents an intermediate super-loading scenario; and the no doublets scenario serves as a best-case baseline scenario to evaluate performance of the demultiplexing tools.

Single-cell RNA sequencing of ovarian tumors

De-identified HGSOC samples were processed after cryopreservation in liquid nitrogen where tissue chunks were stored in RPMI media with 10% fetal bovine serum and 10% DMSO. Samples were thawed and dissociated to single cells using the Miltenyi Human Tumor Dissociation Kit and the GentleMACS dissociator. Samples were incubated on the GentleMACS at 37°C for 1 hour with the setting of 1,865 rounds per run. A 70 µm MACS smart strainer was used to deplete cell doublets before loading onto the 10x Genomics Chromium Controller. Library preparation was performed using the 10x Genomics 3’ Gene Expression Library Prep v3 and libraries were sequenced on an Illumina NovaSeq instrument.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data access

Raw and processed sequencing data generated in this study (HGSOC dataset) are available from the Database of Genotypes and Phenotypes (dbGaP) (raw data consisting of FASTQ files, accession phs002262.v1.p1) and Gene Expression Omnibus (GEO) (processed data files containing gene count tables, accession GSE158937). The lung adenocarcinoma dataset was previously published by [36], and is available from the European Genome-phenome Archive (EGA) (EGAD00001005054).

Code access

All code scripts to reproduce the benchmarking evaluations, generate figures in the manuscript, and run the Snakemake workflow are freely accessible from GitHub at https://github.com/lmweber/snp-dmx-cancer. All tools used within the benchmarking evaluations and workflow are freely available, as described in Methods.

Acknowledgments

We thank Yuanhua Huang for assistance with running Vireo and cellSNP; Davis McCarthy for advice regarding Vireo; and attendees from the Stephanie Hicks and Kasper Hansen joint lab meetings at Johns Hopkins University for helpful feedback and discussions. We thank Hae-Ock Lee and Myung-Ju Ahn of the Samsung Medical Center for providing access to the lung adenocarcinoma dataset. Research reported in this publication utilized the Biorepository and Molecular Pathology Shared Resource and the High-Throughput Genomics Shared Resource at the Huntsman Cancer Institute at University of Utah and was supported by NIH/NCI award P30 CA042014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author contributions

LMW: Software, Formal analysis, Investigation, Data curation, Writing - Original Draft, Visualization
AAH: Software
PFH: Software, Writing - Review & Editing
KCB: Investigation
JG: Investigation, Resources, Writing - Review & Editing
JAD: Resources, Writing - Review & Editing, Funding acquisition
CSG: Conceptualization, Resources, Writing - Review & Editing, Funding acquisition
SCH: Conceptualization, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Ethics approval and consent to participate

Ovarian cancer tissue was obtained and studied under written informed consent at the Huntsman Cancer Institute through approved University of Utah Institutional Review Board protocols IRB_00010924 and IRB_00118086. Analysis of human data in this study was also approved by the University of Pennsylvania Institutional Review Board (IRB protocol 832353) and the Johns Hopkins Bloomberg School of Public Health Institutional Review Board (IRB00013099).

Competing interests

The authors declare no conflicts of interest.

Funding

LMW, AHA, KCB, JG, JAD, CSG, and SCH were supported by the National Institutes of Health grant from the National Cancer Institute R01CA237170. JAD is also supported by Huntsman Cancer Foundation and National Institutes of Health grant from the National Cancer Institute P30 CA042014 (to N. Ulrich).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018;19: 562–578.

2. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. Scientific Reports. 2017;7: 39921.

3. Huang Y, McCarthy DJ, Stegle O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. Genome Biology. 2019;20: 273.

4. Xu J, Falconer C, Nguyen Q, Crawford J, McKinnon BD, Mortlock S, et al. Genotype-free demultiplexing of pooled single-cell RNA-seq. Genome Biology. 2019;20: 290.

5. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology. 2018;36: 89–94.

6. Germain P-L, Sonrel A, Robinson MD. pipeComp, a general framework for the evaluation of computational pipelines, reveals performant single cell RNA-seq preprocessing tools. Genome Biology. 2020;21.

7. Bais AS, Kostka D. scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics. 2019;36: 1150–1158.

8. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Systems. 2019;8: 281–291.e9.

9. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Systems. 2019;8: 329–337.e4.

10. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research. 2016;5: 2122.

11. Farouni R, Djambazian H, Ferri LE, Ragoussis J, Najafabadi HS. Model-based analysis of sample index hopping reveals its widespread artifacts in multiplexed single-cell RNA-sequencing. Nature Communications. 2020;11: 2704.

12. Griffiths JA, Richard AC, Bach K, Lun ATL, Marioni JC. Detection and removal of barcode swapping in single-cell RNA-seq data. Nature Communications. 2018;9: 2667.

13. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics. 2018;19: 332.

14. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ. Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv (https://www.biorxiv.org/content/101101/125724v1) (doi: https://doi.org/101101/125724). 2017.

15. Illumina. Effects of index misassignment on multiplexing and downstream analysis. White Paper (https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf). 2018.

16. McGinnis CS, Patterson DM, Winkler J, Conrad DN, Hein MY, Srivastava V, et al. MULTI-seq: sample

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nature Methods. 2019;16: 619–626.

17. Stoeckius M, Zheng S, Houck-Loomis B, Hao S, Yeung BZ, Mauck WM 3rd, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biology. 2018;19: 224.

18. Xin H, Lian Q, Jiang Y, Luo J, Wang X, Erb C, et al. GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. Genome Biology. 2020;21.

19. Heaton H, Talman AM, Knights A, Imaz M, Gaffney DJ, Durbin R, et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. Nature Methods. 2020;17: 615–620.

20. Zhang F, Kang HM. popscle: A suite of population scale analysis tools for single-cell genomics data (freemuxlet). Software package (https://github.com/statgen/popscle). 2020.

21. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74.

22. Spencer DH, Zhang B, Pfeifer J. Chapter 8 - Single Nucleotide Variant Detection Using Next Generation Sequencing. In: Kulkarni S, Pfeifer J, editors. Clinical Genomics. 2015. pp. 109–127.

23. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578: 82–93.

24. Fancello L, Gandini S, Pelicci PG, Mazzarella L. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. Journal for Immunotherapy of Cancer. 2019;7: 183.

25. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Medicine. 2017;9: 34.

26. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001;29: 308–311.

27. Hafemeister C, Satija R. Cost Per Cell: Multiplexing cost calculator. Website (https://satijalab.org/costpercell). 2018.

28. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. eLife. 2017;6: 1–30.

29. Taylor DM, Aronow BJ, Tan K, Bernt K, Salomonis N, Greene CS, et al. The Pediatric Cell Atlas: Defining the Growth Phase of Human Development at Single-Cell Resolution. Developmental Cell. 2019;49: 10–29.

30. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nature Genetics. 2013;45: 1127–1133.

31. Köster J, Rahmann S. Snakemake — a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522.

32. 10x Genomics. Cell Ranger. Software (https://support10xgenomics.com/single-cell-gene-expression/software/overview/welcome). 2020.

33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.06.371963; this version posted November 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

34. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27: 2987–2993.

35. Huang Y. cellSNP. Software package (https://github.com/single-cell-genetics/cellSNP). 2020.

36. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nature Communications. 2020;11: 2285.

37. Srivastava A, Malik L, Smith T, Sudbery I, Patro R. Alevin efficiently estimates accurate gene abundances from dscRNA-seq data. Genome Biology. 2019;20: 65.

38. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017;8: 14049.

39. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.